DOCUMENT RESUME

ED 209 326 TM 810 868

AUTHOR Modu, Christopher C.
TITLE Effects of Different Methods of Weighting Subscores on the Composite-Score Ranking of Examinees.
INSTITUTION College Entrance Examination Board, New York, N.Y.
REPORT NO CB-R-81-2
PUB DATE 81
NOTE 17p.

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Advanced Placement Programs; *Correlation; Equated Scores; Higher Education; *Measurement Techniques; Scoring; *Scoring Formulas; *Statistical Analysis; *Weighted Scores
IDENTIFIERS Advanced Placement Examinations (CEEB); *Composite Scores; *Ranking

ABSTRACT

The effects of applying different methods of determining different sets of subscore weights on the composite score ranking of examinees were investigated. Four sets of subscore weights were applied to each of three examination results. The scores were from Advanced Placement (AP) Examinations in History of Art, Spanish Language, and Chemistry. One set, operational weights, was determined in advance of the test administration. The other three sets, simplified weights, Wilks's Method C weights, and standardized scores, were generated after the tests were scored. Each set of weights was intended to reflect the prescribed proportional contribution of each subscore. The appeal for the set generated in advance derives from its time and cost saving considerations, since final results showed little difference which weighting procedure was used. If any weighting procedure is to be avoided in order to maintain consistency with operational weighting, it is the one involving the transformation of subscores to standardized scores. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED209326

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

College Board Report

No. 81-2

Effects of Different Methods of Weighting Subscores on the Composite-Score Ranking of Examinees

Christopher C. Modu

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

P. K. Handel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Effects of Different Methods of Weighting Subscores on the Composite-Score Ranking of Examinees

Christopher C. Modu

College Board Report No. 81-2

College Entrance Examination Board, New York, 1981

Christopher C. Modu is a staff member of the Educational Testing Service, Princeton, New Jersey.

Grateful acknowledgment is due to Sandy Richards for writing the computer programs used for this study, to Edwin O. Blew for his assistance in generating some of the computer outputs, and to Ikuko Nutkowitz for organizing some aspects of the data.

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,500 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publication Orders, Box 2815, Princeton, New Jersey 08541. The price is $4.00.

Copyright © 1981 by College Entrance Examination Board.
All rights reserved.
Printed in the United States of America.

CONTENTS

Abstract 1

Introduction 2

Method 2

Results 5

Conclusion 9

References 11

ABSTRACT

The effects of applying different methods of determining different sets of subscore weights on the composite-score ranking of examinees were investigated. Four sets of subscore weights were applied to each of three separate examination results. One set was determined in advance of the test administration, the other three sets were generated after the tests were scored. Each set of weights was intended to reflect the prescribed proportional contribution of each subscore. Since the results showed that it made little difference which weighting procedure was used, the appeal for the set generated in advance derives from its time- and cost-saving considerations.

INTRODUCTION

This study investigates the effects of different methods of determining subscore weights on the composite-score ranking of examination candidates. Stalnaker (1938) had demonstrated, in an earlier study, that if the scores on the questions of an examination are highly interrelated, then the choice of weights for the part scores becomes a less important issue in the stability of the composite-score ranking of the candidates. In that study, Stalnaker determined the effect of using different sets of subscore weights by finding the relationship between linear combinations of weighted and unweighted subscores for five mathematics examinations, each with 11 or more scorable units, and for an English examination of six questions. He obtained correlations of .98 to .99 for the two sets of composite scores on the five mathematics examinations. In English, the correlation between the two sets of total scores obtained by applying the subscore weights: 14, 24, 32, 15, 9, and 34 versus the simpler weights: 1, 2, 2, 1, 1, and 3 to the six questions was found to be .997. Also, the correlation between the unweighted English scores and the scores weighted by the simpler system was found to be .97. Therefore, a replication of Stalnaker's results in this study should eliminate the unnecessary concern over whether operational subscore weights should be rounded to whole numbers or carried to several decimal places, or over how accurately somewhat dissimilar operational weights reflect the intended proportional contribution of each subscore to the composite-score distribution.

The present study is important for two reasons. First, it will show whether simplified weights (which will save time in the hand computation of composite scores for candidates with irregularity reports and those used in quality control of reported scores) can be used without any noticeable change in the rank-ordering of the candidates. Second, it will show whether weights now used operationally have effects similar to those associated with more optimal weights such as the one referred to later in this report as Wilks's Method C (Wilks, 1938, pp. 35-39) which makes use of correlations among the subscores and ensures that a given subscore distribution makes the desired proportional contribution to the composite-score variance.

METHOD

Scores from three Advanced Placement (AP) Examinations in History of Art, Spanish Language, and Chemistry have been used for the study, because the three examinations are considered to be representative of the various subject areas covered in that program. Specifically, the History of Art examination, which has no multiple-choice component, comprises 15 essay-type questions grouped into four subscores labeled EPT1-4 (i.e., Essay Part Scores, 1-4); Spanish Language contains a 90-item multiple-choice section and an essay section of five part scores; and Chemistry contains an 80-item multiple-choice section and an essay section of six questions grouped into four part scores. A total of 451 candidates in History of Art, and representative samples of 1,183 candidates for Spanish Language, and 1,684 candidates for Chemistry, were selected from the May 1977 operational administration of the Advanced Placement Examinations for the study.

Four composite-score distributions were obtained for each examination by the following methods in which different sets of weights were derived for the part scores.

Methods Based on Maximum Possible Score

(i) The operational weights are those in which the subscore weights were determined, as in the May 1977 administration, by expressing the maximum possible subscores as prescribed percentages of the maximum possible composite score. These percentages are usually prescribed by the Test Development Committees.

(ii) The simplified weights are those in which such operational subscore weights as 0.6, 0.675, 1.467, 2.222, and 3.6, determined as described above, were rounded to simple rational numbers like 1, 3/4, 3/2, 2, and 4, respectively.

Method Based on Correlations and Standard Deviations

(iii) Wilks's Method C weights are derived to ensure that a given subscore makes the desired proportional contribution to the composite-score variance. In this regard, it must be noted that neither the common practice of multiplying standard scores by predetermined or a priori weights, nor that of calculating the maximum possible subscore as a given percentage of maximum total score, yields a total-score distribution in which the component subscores make prescribed proportional contributions to the total-score variance. Rather, the desired results are accomplished by applying a mathematical solution in which the derived weights are used to increase or decrease the scattering effects of given subtests on the variation of the composite scores in proportion to the different a priori importance attached to the subtests. Method C is most appropriate when the influence or weight of a question in the total score is a function of the differentiating power of the question and of its relation to other questions on the same examination. A disadvantage of this method is that, unlike the first two methods, the subscore weights can only be determined from the summary statistics and correlations to be obtained after the scoring is completed.

Method Based on Standard Deviations Only

(iv) Standardized scores are generated from weights expressed as linear conversion parameters for transforming each of the subscores for an examination to a specified mean and standard deviation (i.e., M = 90, SD = 20 for History of Art; M = 140, SD = 40 for Spanish Language; and M = 90, SD = 30 for Chemistry). Thus, if A and B are the slope and intercept of the conversion line which will transform the EPTl subscore distribution for History of Art to the specified mean and standard deviation, then the weight (or conversion parameters) to be applied to each obtained score, x_i, of EPTl becomes .35 (A x_i + B), where .35 is the prescribed proportional contribution for that subscore to the total score; and similarly for other subscores. As in Wilks's Method C, the weight (or conversion parameters) for each subscore cannot be determined in advance until all scoring is completed.

The part scores for each examination, the maximum possible unweighted score for each part score, the effective weight for the part score as prescribed by the committee of examiners, and the operational or experimental weights to be applied to each part score are listed in Table 1, which follows, for each of the three examinations. Thus, for each examination, four total scores have been computed for each candidate, one based on the operational weights and the remaining three on the experimental weights, namely: the simplified and the Wilks's Method C sets of weights, as well as the set of weights expressed as linear conversion parameters to be applied to different subscores.

The "percent of maximum possible score" column to the right of the operational and the experimental sets of subscore weights in Table 1 presents the maximum possible subscore for a given variable as a percentage of the maximum possible composite score. The maximum possible subscores and composite score are obtained by applying the weights in the preceding column to the corresponding unweighted maximum possible subscores in the second column of the tables.

TABLE 1. Operational and Experimental Weights for Subscores, and Each Maximum Subscore Expressed as a Percentage of the Maximum Possible Composite Score

Subscore	Unwt'd Max. Possible Score	Prescribed Proportion (p) of Subscore as % of Total Score	Operational Weight	% of Max. Poss. Score	Simpler Weight	% of Max. Poss. Score	Wilks's[a] Weight	% of Max. Poss. Score	Slope A and Intercept B' of Raw to Standardized-Score Conversion Line[b] pA	pB	Max. Poss. Standardized Score	Max. Std. Score as % of Max. Comp. Score	
History of Art													
EPT1	28	35.0	2.0	35.0	2.0	36.8	2.0983	38.8	2.0983	-4.3372	54.42	34.8	
EPT2	24	15.0	1.0	15.0	1.0	15.8	1.0739	17.0	1.0120	-0.8558	23.43	15.0	
EPT3	18	25.0	2.222	25.0	2.0	23.7	2.0368	24.2	1.6962	10.6761	41.21	26.3	
EPT4	18	25.0	2.222	25.0	2.0	23.7	1.6852	20.0	1.3941	12.2343	37.33	23.9	
Maximum Possible Composite Score:			160		152		152				156		
Spanish Language													
EPT1	15	6.0	0.900	6.0	1.000	6.25	1.1388	7.0	0.7897	0.9190	12.76	5.9	
EPT2	15	24.0	3.600	24.0	4.000	25.0	3.7470	23.0	3.1103	5.5228	52.18	24.0	
EPT3	20	6.0	0.675	6.0	0.750	6.25	1.0143	8.3	0.7775	-3.6187	11.93	5.5	57.1
EPT4	40	12.0	0.675	12.0	.750	12.5	0.5458	8.9	0.4892	3.5446	23.11	10.6	
EPT5	15	12.0	1.800	12.0	.00	12.5	1.4080	8.6	1.2122	6.0766	24.26	11.1	
Objective	90	40.0	1.000	40.0	.000	37.5	1.2041	44.2	0.9633	6.5190	93.22	42.9	
Maximum Possible Composite Score:			225		240		245				217		
Chemistry													
EPT1	15	13.75	1.467	13.75	1.5	13.1	1.1317	10.1	0.9737	4.8146	19.42	12.0	
EPT2	15	13.75	1.467	13.75	1.5	13.1	1.2544	11.2	0.9623	4.4108	18.84	11.6	
EPT3	15	8.25	0.880	8.25	1.0	8.8	0.8764	7.9	0.7032	1.0298	11.58	7.	50.0
EPT4	21	19.25	1.467	19.25	1.5	18.4	1.5612	19.6	1.3445	2.8667	31.10	19.2	
Objective	80	45.0	0.900	45.0	1.0	46.6	1.0712	51.2	0.9038	8.5750	80.88	50.0	
Maximum Possible Composite Score:			160		172		167				162		

[a]Determined so that each subscore contributes the prescribed proportional weight to composite-score variance.

[b]$A' = pA$; $B' = pB$, where A and B are the slope and intercept of the conversion line which transforms each subscore to the specified mean and standard deviation for an examination, and p is the prescribed proportion of a given subscore as percent of total or composite score.

RESULTS

The intercorrelations among the four essay part scores for History of Art range from .273 to .613. Among the objective and five essay subscores for Spanish Language they range from .397 to .860, and from .433 to .714 among the objective and four essay subscores for Chemistry.

The cut scores (for the five AP grade levels) determined from the composite-score distributions for the operational administration are provided in Table 2. Their equivalent cut scores in the composite-score distributions generated by applying the three experimental sets of weights to the subscores for the same group of candidates for each examination are also given in the last three columns of Table 2. Cut scores at the same percentile rank in each of the four composite-score distributions for each examination are considered to be equivalent. Thus, equivalent cut scores are assumed to represent the same ability level regardless of the weighting procedure used in obtaining the different sets of composite scores from which they are derived.

TABLE 2. Equivalent Cut Scores in the Composite-Score Distributions Based on Different Sets of Subscore Weights

Grade	Operational Wts.	Simplified Wts.	Wilks's Method C Wts.	Standardized Score Wts.
History of Art				
5	102-160	98-152	98-152	109-156
4	91-101	87- 97	88- 97	99-108
3	70- 90	68- 86	69- 87	82- 98
2	56- 69	54- 67	56- 68	70- 81
1	0- 55	0- 53	0- 55	0- 69
Spanish Language				
5	191-225	204-240	206-245	187-217
4	146-190	155-203	159-205	148-186
3	116-145	124-154	128-158	122-147
2	83-115	89-123	92-127	93-121
1	0- 82	0- 88	0- 91	0- 92
Chemistry				
5	111-160	119-172	115-167	118-162
4	92-110	99-118	94-114	101-117
3	61- 91	65- 98	63- 93	75-100
2	42- 60	45- 64	44- 62	58- 74
1	0- 41	0- 44	0- 43	0- 57

The intercorrelations among the four composite-score distributions based on the different weighting methods are presented below in Table 3 along with the corresponding intercorrelations among the four AP grade distributions derived by applying the cut scores in Table 2 to the respective composite-score distributions.

TABLE 3. Intercorrelations for Composite Scores or Grades Based on Different Sets of Weights

	Composite Scores				AP Grades			
History of Art: (N = 481)	1	2	3	4	1	2	3	4
1. Operational Wts.	-				-			
2. Simplified Wts.	.9994	-			.9907	-		
3. Wilks's Wts.	.9964	.9982	-		.9698	.9757	-	
4. Standardized Wts.	.9922	.9953	.9986	-	.9505	.9578	.9763	-
Spanish Language: (N = 1,183)								
1. Operational Wts.	-				-			
2. Simplified Wts.	.9999	-			.9939	-		
3. Wilks's Wts.	.9987	.9981	-		.9813	.9802	-	
4. Standardized Wts.	.9993	.9988	.9998	-	.9871	.9841	.9917	-
Chemistry: (N = 1,684)								
1. Operational Wts.	-				-			
2. Simplified Wts.	.9998	-			.9928	-		
3. Wilks's Wts.	.9983	.9988	-		.9808	.9837	-	
4. Standardized Wts.	.9978	.9984	.9998	-	.9772	.9822	.9925	-

The correlations among the four sets of composite scores derived for each examination by using different subscore weights range from .9922 to .9999. Corresponding correlations among sets of AP grades range from .9505 to .9939. These correlations are so high that the use of any of the four weighting procedures would have made little difference to the final AP grades of the candidates. The slightly lower correlations among the sets of AP grades relative to the composite-score correlations may have resulted from two factors: the more restricted range (1-5) of the AP grade scale; and the slight shifts in the percentages at corresponding AP grade distributions for the different weighting procedures due to the rounding of the equivalent cut scores to the nearest composite-score integer values.

The number and percentage of candidates at each AP grade level are presented in Table 4 for the operational and the three experimental weighting procedures. Each grade

distribution is obtained by applying the cut scores displayed in Table 2 to the composite-score distribution for the respective weighting procedure. As Table 4 clearly indicates, the percentage of candidates at each grade level is fairly stable across different weighting procedures. Failure to achieve identical grade distributions is attributable to the practice of rounding composite scores to integer values before applying the equated cut scores to the distributions for each examination. The percentages in the top three grades, 3-5, for which credit or advanced placement is generally awarded to candidates, are as follows across the four weighting procedures: 70.7, 70.1, 69.6, and 70.7 for History of Art; 69.2, 69.0, 68.7, and 69.1 for Spanish Language; and 71.7, 72.3, 71.8, and 71.3 for Chemistry

TABLE 4. Comparative AP Grade Distributions Under Different Weighting Procedures

AP Grade	Operational Wts. N	Operational Wts. (% At)	Simplified Wts. N	Simplified Wts. (% At)	Wilks's Wts. N	Wilks's Wts. (% At)	Std. Score Wts. N	Std. Score Wts. (% At)
History of Art								
5	62	(12.9)	60	(12.5)	62	(12.9)	59	(12.3)
4	75	(15.6)	78	(16.2)	65	(13.5)	79	(16.4)
3	203	(42.2)	199	(41.4)	208	(43.2)	202	(42.0)
2	98	(20.4)	104	(21.6)	104	(21.6)	100	(20.8)
1	43	(8.9)	40	(8.3)	42	(8.7)	41	(8.5)
	481		481		481		481	
Spanish Language								
5	126	(10.7)	127	(10.7)	127	(10.7)	123	(10.4)
4	396	(33.5)	403	(34.1)	400	(33.8)	400	(33.8)
3	296	(25.0)	286	(24.2)	286	(24.2)	295	(24.9)
2	236	(19.9)	239	(20.2)	243	(20.5)	238	(20.1)
1	129	(10.9)	128	(10.8)	127	(10.7)	127	(10.7)
	1,183		1,183		1,183		1,183	
Chemistry								
5	256	(15.2)	254	(15.1)	250	(14.8)	247	(14.7)
4	373	(22.1)	363	(21.6)	390	(23.2)	374	(22.2)
3	579	(34.4)	600	(35.6)	570	(33.8)	580	(34.4)
2	274	(16.3)	271	(16.1)	275	(16.3)	288	(17.1)
1	202	(12.0)	196	(11.6)	199	(11.8)	195	(11.6)
	1,684		1,684		1,684		1,684	

The net shifts in the grade distributions reported in Table 4 do not, however, reveal the actual changes from one grade level to another between pairs of weighting procedures. These are best illustrated by cross-tabulations which will show whether one weighting procedure compared to another resulted in shifts of one, two, three, or four grade levels. Thus, six cross-tabulations of grades were generated for pairs of the four weighting procedures for each examination. In none of the 18 cross-tabulations for the three examinations was a shift greater than one grade level observed. No shifts would have placed all the entries in the main diagonal of each cross-tabulation.

Two cross-tabulations of the number of cases at each grade--one with the lowest and the other with the highest Pearson's coefficient of correlation (see Table 3) for grades from pairs of weighting procedures--are presented below. The other cross-tabulations lie between the two extremes.

History of Art:

		Simplified Wts.				
	Grade	1	2	3	4	5
Operational Wts.	1	40	3			
	2		98			
	3		3	199	1	
	4				74	1
	5				3	59

R = .9907

		Std. Score Wts.				
	Grade	1	2	3	4	5
Operational Wts.	1	40	3			
	2	1	87	10		
	3		10	182	11	
	4			10	60	5
	5				8	54

R = .9505

Spanish Language:

		Simplified Wts.				
	Grade	1	2	3	4	5
Operational Wts.	1	127	2			
	2	1	233	2		
	3		4	284	8	
	4				394	2
	5				1	125

R = .9939

		Wilks's Wts.				
	Grade	1	2	3	4	5
Simplified Wts.	1	125	3			
	2	2	227	10		
	3		13	260	13	
	4			16	383	4
	5				4	123

R = .9802

Chemistry:

Simplified Wts.

Operational Wts. \ Grade	1	2	3	4	5
1	196	6			
2		265	9		
3			577	2	
4			14	358	
5				3	253

R = .9928

Std. Score Wts.

Operational Wts. \ Grade	1	2	3	4	5
1	189	13			
2	6	253	15		
3		22	542	15	
4			23	346	4
5				13	243

R = .9772

The above cross-tabulations indicate that the most stable results were obtained between the operational and the simplified sets of weights. In no case was a shift of more than one grade level observed from one weighting procedure to another. A closer scrutiny of the cross-tabulations for the operational versus the simplified weights shows that, for the total of 11 grade changes in History of Art, the use of simplified weights rather than the operational would have resulted in one grade level higher for 5 candidates but one grade level lower for 6 out of 491 candidates. Similar comparisons produce a higher grade for 14, but a lower grade for 6, out of 1,183 Spanish Language examination candidates; and a higher grade for 18, but a lower grade for 17, out of 1,684 Chemistry examination candidates. The other three cross-tabulations suggest that if any one of the weighting procedures is to be avoided in order to maintain consistency with the results from the operational weights, it is that involving the transformation of subscores to standardized scores. In no case, however, was there a very high percentage of grades affected.

CONCLUSION

The effect of applying different methods of determining different sets of subscore weights on the composite-score ranking of candidates for three Advanced Placement Examinations was examined in this study. Four sets of subscore weights were applied to each examination. One set, used for the operational administrations, was calculated to yield maximum possible subscores which are prescribed percentages of the maximum possible composite score. The other three were experimental sets of weights in which (a) the operational weights were rounded to simple rational numbers, (b) a given subscore makes a prescribed proportional contribution to the composite-score variance, and (c) the subscore standard deviations are proportional to the prescribed contribution of each subscore to the composite score.

The results of the study indicate that paired sets of composite scores have correlations of .99 or higher for all four methods of weighting procedures. Paired sets of final AP grades determined from the composite-score sets through equated cut scores also produced correlations between .96 and .99. In no case was a shift of more than one grade level observed between the results of any pair of weighting procedures.

Since it makes little difference to the final results which weighting procedure is used, and in view of the fact that the operational and simplified sets of weights can be determined well in advance of the scoring process, it is recommended that either of these two methods be used for AP Examinations. If any weighting procedure is to be avoided in

order to maintain consistency with operational weighting, it is the one involving the transformation of subscores to standardized scores. However, the fact that the essay papers for Advanced Placement Examinations are scored in such a manner as to yield a full range of possible scores for each question may have considerably minimized the value of weighting the subscores as prescribed proportions of their standard deviations. This procedure might well be recommended for use in situations where the scoring procedure tends to bunch most scores in a restricted range of the score scale.

Also, moderate or high intercorrelations among the subscores do not appear to be a satisfactory basis for recommending the selection of one weighting procedure over another, considering that consistent results were obtained in the study for all three examinations despite the fact that the intercorrelations among four History of Art subscores range from a low of .273 to a high of .613, whereas those for Spanish Language and Chemistry range from .397 to .860 and .433 to .714, respectively.

REFERENCES

Stalnaker, John M. "Weighting Questions in the Essay-Type Examination," Journal of Educational Psychology, 28:7 (October 1938): 481-490.

Wilks, S.S. "Weighting Systems for Linear Functions of Correlated Variables When There Is No Dependent Variable," Psychometrika, 3:1 (March 1938): 23-40.